专家论坛

微生物组学大数据分析方法、挑战与机遇

盛华芳,周宏伟
南方医科大学公共卫生与热带医学学院环境卫生系,广东 广州 510515

摘要: 微生物组学是新兴学科,与肠道、代谢、生殖、神经等大量慢性疾病相关。通过测序分析微生物组,主要包括16S rRNA和宏基因组两大技术。16S rRNA数据分析主要包括序列处理、样品多样性分析及统计分析3个步骤。宏基因组数据分析主要包括序列处理、分类、注释及统计分析4个环节。随着测序技术的升级,测序成本将逐步降低,而大数据分析将成为核心内容。数据的标准化和可积累性、通过数据建模和预测疾病的发生发展是未来应用的基础,数据知识产权保护以及数据本身价值的开发与保护价值将日益显现,培养和基于培养的功能验证将是未来的重点之一。人体微生物组学将阐述并调整人与微生物组之间的关系,此领域相关研究有巨大的发展空间。
关键词:微生物组学;大数据分析;宏基因组学;16S

Methods, challenges and opportunities for big data analyses of microbiome

SHENG Huafang, ZHOU Hongwei
Department of Environmental Health, School of Public Health and Tropical Medicine, Southern Medical University, 510515 China

Abstract: Microbiome is a novel research field related with a variety of chronic inflamatory diseases. Technically, there are two major approaches to analysis of microbiome: metataxonome by sequencing the 16S rRNA variable tags, and metagenome by shot-gun sequencing of the total microbial (mainly bacterial) genome mixture. The 16S rRNA sequencing analyses pipeline includes sequence quality control, diversity analyses, taxonomy and statistics; metagenome analyses further includes gene annotation and functional analyses. With the development of the sequencing techniques, the cost of sequencing will decrease, and big data analyses will become the central task. Data standardization, accumulation, modeling and disease prediction are crucial for future exploit of these data. Meanwhile, the information property in these data, and the functional verification with culture-dependent and culture-independent experiments remain the focus in future research. Studies of human microbiome will bring a better understanding of the relations between the human body and the microbiome, especially in the context of disease diagnosis and therapy, which promise rich research opportunities.
Key words: microbiome; big data analyses; metagenome; 16S

“微生物群落”是地球上生命基本元素(C、N和S等)进行生物地球化学循环的主要驱动力,与人类健康、环境保护以及工农业生产等密切相关。近十年来,随着高通量测序的广泛应用,“微生物组学”成为新兴概念和热点。微生物组与不同的生存环境结合,诞生人体微生物组,宿主相关微生物组,一般环境微生物组,建筑环境微生物组,地球微生物组,医院环境微生物组等大量新兴的研究方向。

长期以来,研究方法一直是微生物群落研究的瓶颈。如,群落结构的阐述,即准确描述一定空间范围内的物种数量,并定量各物种的丰度,这是所有生态学研究的基本内容。然而,对微生物研究者而言,实现这一基本要求却绝非易事。这种困难主要源于微生物群落的如下几点特征。

(1)“微小”:宏观生物可以肉眼或者镜下观察其形态学分类特征并计数;而微生物即便在显微镜下也难以区分,形态差异特征少,因而不能直接观察种属并计数。

(2)“复杂”:微生物很少以纯种存在;但微生物群落含有极高的多样性。1 g土壤中可能含有数千到数万个不同种属的微生物。

(3)“稠密”:1 g土壤,1滴流水,都可能含有数以十亿计的微生物,并且它们常常来自成千上万的种属。

(4)“不均”:不同种属微生物在群落中的丰度差异极大。这种不均匀的分布特征造成优势种、非优势种以及稀有种的计数难以同时进行。

面对如此庞大复杂的微生物生态系统,微生物组学要准确理解样品中的微生物种类,多度及其功能,并将

收稿日期:2015-04-10
基金项目:国家自然基金(31322003,31270152)
Supported by National Natural Science Foundation of China (31322003, 31270152).
作者简介:盛华芳,硕士,助理实验师,电话:020-62789126,E-mail: shenghuafang0727@163.com
通信作者:周宏伟,南方医科大学公共卫生与热带医学环境卫生系主任、教授、博士生导师,器官衰竭防治国家重点实验室PI。国家优秀青年基金获得者,教育部新世纪优秀人才、广东省“千百十人才工程”国家级培养对象。E-mail: biodegradation@gmail.com

其与时间、空间、理化因素，宿主疾病状态等进行关联，从而探求微生物与微生物之间，微生物与宿主之间，以及微生物与环境之间的相互关系。因此，需要恰当的技术，在广度和精度这两个略显矛盾的角度，同时获得理想的数据。

自2006年，随着新一代高通量测序技术的成熟，不仅在人类基因组学领域带来了翻天覆地的变化，对微生物组学的研究产生了革命性的影响。当前，以16S rRNA高通量测序为基本手段，宏基因组鸟枪法测序、宏转录组、宏蛋白组、宏代谢组等组学领域产生了大量的新技术，共同促进了微生物组学的快速进步。

1 微生物组大数据分析的方法和流程

16S的测序是近年来微生物生态领域最核心、最重大的突破。通过454、Illumina等第2代测序仪高通量测定16S可变区序列，第1次让人们在可行的成本下，获得全面、系统、结构化的群落结构信息[1-2]。美国Woods Hole海洋研究实验室的Mitchell Sogin课题组于2006年首次报道了通过焦磷酸测序技术，测定海洋沉积物样品的16S rRNA基因V6可变区，人类第1次在基本足够的测序深度下，清晰地展示了环境样品中微生物的组成，发现了高度的多样性。

与所有传统的微生物组学研究方法相比，该方法具有显著的优越性。该方法通过测定16S短片段序列，经生物信息学分析可以获得系统分类信息，从而可以明确定性其分类单元，不同实验间数据完全是可比较、可积累的。该方法通量显著提高，1次测定40~100万条序列，通过条码技术可以对每个样品测定数千到数万条短序列，从而可以获得广泛的、系统的结构信息。由于测序深度大，在多个数量级范围内可以进行定量。该方法的诞生对微生物组学的研究产生了巨大的影响，尤其对人体共生微生物领域最为活跃。例如，肥胖与部分肠道菌群间的相关性研究[3]；人体不同部位的菌群结构的首次阐明[4]；抗生素对肠道微生物群落产生的显著影响[5]等。在环境中，该技术首次在海洋沉积物中发现存在极其丰富、多样化的微生物群落。该方法让人们得以比较大空间尺度下土壤微生物群落结构的差异及其主要的影响因素（如pH）[6-8]。

基于16S的分析可称为宏分类组技术（metataxonome）。16S的数据分析，其一般流程包括：序列提取、质控、相似序列聚类成OTU，种属分类，alpha以及beta多样性分析，以及进一步的统计分析。其中每一步都有关键之处，并正处于方法学前沿领域。

OTU聚类是16S序列分析的关键问题之一。在经典的分层聚类算法中，其运算量和所需的内存容量，均随着序列数量的增加呈几合级数增加。因此，贪婪算法成为目前该领域的主流。同时，也有不少研究者开发不基于序列比对的聚类算法。但是，由于序列相似性算法的不同，聚类中距离的传递问题，以及参考序列数据库的不足，该领域仍然存在运算效率和准确性问题。目前，与参比库比对的Open-reference算法[9]以及UPARSE是运用较为广泛的技术。

在完成聚类后，种属的分类仍然存在许多问题。目前，该领域主要通过与16S数据库比对，选取相似性高的参比序列的分类结果。但是，参比数据库本身，目前存在不少问题，例如目前应用最为广泛的Greengenes数据库[10]，其中不少序列存在重复或者错误的分类结果。

UniFrac距离的计算，是beta多样性分析的关键工具。UniFrac距离是美国科罗达罗大学Rob Knight课题组创建的一种基于序列之间相似度，计算样品之间总的菌群距离的算法，有加权和不加权两种，在分析微生物群落相似性中均具有重要作用[11]。基于UniFrac的工作基础，Rob Knight课题组进一步开发了微生物群落以及微生物生态分析的主流工具体系Quantitiative Insight Into the Microbial Ecology（QIIME）。该平台是一个流程的整合，已经在全球分析微生物组学科中广泛应用[12]。

与之对应，Patrick Schloss开发了Mothur[13]，该平台基于最初的序列聚类工具DOTUR而来。该平台和QIIME竞争，在许多地方有相似之处。二者之间的区别是，QIIME更为开放，系统整合能力更强，尊重方法的原创者，应用者更多一些，而Mothur则全部经作者改写，相对封闭。核糖体数据库RDP database课题组，也同样开发了针对二代测序数据的群落分析工具[14]。除此之外，MG-RAST是一个综合性的在线数据分析平台[15]。使用者只需要将自己的测序数据投递到该网站，即可点击不同的宏基因组分析命令，完成数据分析。欧洲MetaHIT以及其它小组也开发了一些微生物群落的分析工具，但应用面不及上述几个平台。

需要指出的是，除了16S外，人们还开发了一些针对特定功能基因的靶向测序技术，从而检测其功能多样性。其分析流程大体与16S相似，但需要特定的数据库加以比对分析[16]。

宏基因组技术（metagenome），又称为元基因组技术，是在16S分析的基础上，通过宏基因组的鸟枪法高通量测序，能够同时获得菌群的分类信息以及功能基因的数据。并且该技术未经PCR扩增，因此PCR导致的偏差较少（测序建库时还会有部分PCR的影响）。因为微生物群落中不同微生物的多度差异极大，欲获得足够的定量信息，需要测试大量的数据。根据不同的需求，单个样品宏基因组测序的数据量，在Giga以上1~2个数量级水平。如此巨大的数据量，无论是测试成本，还是

分析所需要消耗的机时,都相当可观。因此,人们通常在16S测试的基础上,挑选少量目标样品,测试其全基因组。当前,宏基因组数据的分析,通常包括如下步骤:序列质控;将获得的高质序列组装(或者不经组装,直接与参比数据库比对);将组装后的序列与现有的微生物基因数据比对,并将比对上的序列进行门、纲、目、科、属、种的分类和丰度统计;进行样品间物种多样性的比较,如PCA分析、聚类分析、筛选与样品分组显著相关因子;进行基因组份分析,如前噬菌体预测、可转坐原件、基因预测;通过与KEGG、CAZy、eggNOG数据库比对进行功能注释,分析其中的代谢通路,碳水化合物活性酶、同源性;抗生素耐药组的比对分析等。在宏基因组分析中,针对病毒单独纯化的序列测序,可以获得病毒组数据,对微生物生态的解析,提供了全新的视野。宏基因组测序和16S测序尽管在菌群分布上基本是一致的[17],但分辨效率显著不同。例如,在群落层面,二型糖尿病患者肠道菌群和对照人群并无显著的不同,但是,在宏基因组揭示的功能基因上,两组却呈现显著的差异[18-19]。尽管宏基因组技术非常强大,该技术仍然存在诸多技术瓶颈。其一,大量序列目前尚无法找到匹配的数据库序列,尤其是病毒,大约80%甚至更多的序列无法注释;其二,仅仅通过序列相似度,对功能的注释常常是不准确的,存在大量的误注释;最后,对于大量的微生物基因组,通过宏基因组难以将其进行组装拼接,尤其是对低丰度的菌株。其中前两点缺陷同样适用于宏转录组学。

2 微生物组大数据分析的发展趋势

2.1 数据通量的进一步提高,成本的进一步下降

随着测序、质谱等技术的不断进步,依赖于上述技术的微生物组分析技术将同样不断升级换代。伴随着上述发展,解释微生物组所需要的数据量将不再成为瓶颈。多组学联合应用,将日益成为微生物组大数据分析的常用工具。

2.2 大数据分析成为领域的竞争焦点

当前,数据产生的效率,已经远远高于分析效率的提升。微生物组各类大数据的综合分析,日益成为瓶颈。如何储存,如何积累,如何提取关键信息,如何可视化展示,如何保证数据本身以及分析的可重现性,这一切都成为数据分析的挑战。同时,基于宏基因组数据的网络化分析和展示,是数据分析的新兴方向。

2.3 如何促进数据的标准化和可积累性日益显现

该问题初看不是科学问题,但却是限制我国相关学科发展的重要之处。如此庞大的数据量,如同迷宫一般的分析流程,不断升级的分析工具,许多原有的分析在不断纠错,这些数据的获得都来之不易,如何标准化,如何让后来者可以使用数据,这些对于微生物组学科学问题的解答至关重要。例如,我们发现,经典聚类算法直接导致了OTU的不稳定性,从而导致随着测序深度的变化[1],OTU的组成成员不断发生改变,进而影响到多样性的评估和差异物种的寻找,而该错误在大量的经典文献中都有所体现[20]。有研究者正在探讨,运用iPython notebook工具,将原始数据的分析流程加以保存,从而重现全部的分析过程。

2.4 模型和预测是未来发展的趋势

随着数据量的日益增加,建立微生物生态模型,并预测微生物群落的动态发展,预测相关的生物学效应,是微生物组学研究的重要方向和关键应用。

2.5 数据知识产权保护以及数据本身价值的开发与保护

伴随着微生物组学大数据的不断发展,一个现实的问题是,分析工具网站以及数据储存网站,越来越倾向于要求使用者将原始数据以及研究相关的详细meta-data上传。这一方面使得数据的积累更加可靠。但同时,这种要求使得拥有数据库、掌握数据分析方法的团队,能够比实验者本身更早掌握研究的全面信息。这种大数据挖掘能力,令实验团队面临一定的知识产权损失风险。

2.6 培养和基于培养的功能验证将成为新的瓶颈需求

随着非培养技术提供越来越多的数据,如何验证基于大数据分析提出的假设,如何应用非培养提取出的重要菌株,这一切都需要获得纯培养菌株,在原位和实验室条件下验证其功能。

3 人体微生物组学发展趋势

尽管微生物组学呈现爆炸式发展趋势,该学科尚处在早期阶段,存在巨大的发展空间。作为人体生物医学研究的最后一块处女地(我们不能排除未来在核酸研究等领域存在新发现的可能),整个领域正处于从现象描述,关联分析,到机制研究,模型干预,最终到疾病诊断、预测和治疗的井喷式发展中。

随着数据的积累,人们将日益理解在复杂的人体微生物生态背后,哪些微生物是疾病或健康的驱动者,哪些是沉默者。基于此,我们将能理解何为正常菌群,何为失调菌群。人体微生物组在后天是如何形成、发展、稳定,以及与人体相互作用。从生态学角度,人体微生物组在不同部位形成以及改变的驱动因子是什么?

在应用角度,目前有两个重要的发展方向。其一,运用人体微生物组的组成或者功能基因,预测特定疾病的发生、发展以及结局。过去十多年,人们在运用人体基因组表达谱以及差异基因预测疾病和治疗效果上取得了长足的进步;可以预期,在未来的十年内,人们可以

在微生物组上做出相似的工作。例如,人们发现IBD患者与非患者之间的菌群存在显著差异,可以将其作为非介入式疾病诊断的方法之一[21]。通过对人体微生物组的干预和调整,治疗和预防疾病[22-23]。与人体本身的基因组不同,人体微生物组是可以改变的。人们可以通过窄普抗生素、益生元、药物、饮食调整等策略,改善人体微生物生态。在此类研究中,宿主的选择性、生命早期事件、地区差异,肠道病毒组、人体微生物组之间的传递,以及人体微生物资源开发利用等,都存在大量的微生物生态学问题有待研究。我国传统医学在人体微生物生态方向积累了大量的经验,诸如祛湿、补脾健胃等都可能与人体菌群有关。如何通过现代人体微生物生态学手段,将传统经验进一步提取,澄清其背后的机制,开发药品和保健食品,保护好传统的知识产权,是我国相关科研人员的机遇与责任。

微生物组学研究,目的为回答谁、做什么,和怎么做三大基本问题。如果说过去方法学是核心瓶颈,如今条件则已基本具备。回答谁,人们可以采用16S和宏基因组测序,高通量培养,基因芯片,荧光原位杂交等技术。回答做什么,人们可以采用宏转录组,宏蛋白组,宏代谢组,基因芯片,同位素标记,单细胞测序等手段。回答怎么做,则可以通过上述多组学与环境因子的关联数据挖掘,通过移植实验(例如无菌鼠验证肠道菌群的功能)等手段来解决。随着上述工具的建立,微生物组这一黑盒子正在打开。尽管高通量方法的成本依然偏高,但随着新工具的不断开发,微生物组分析的效率和准确性都将不断提高。如何利用好这些工具,关注微生物生态问题,而不是局限于追求一个完美的技术,值得国内研究者的重视。

参考文献:

[1] Sogin ML, Morrison HG, Huber JA, et al. Microbial diversity in the deep sea and the underexplored "rare biosphere"[J]. Proc Natl Acad Sci USA, 2006, 103(32): 12115-20.

[2] Huber JA, Mark Welch DB, Morrison HG, et al. Microbial population structures in the deep Marine biosphere[J]. Science, 2007, 318(5847): 97-100.

[3] Zhang H, Dibaise JK, Zuccolo A, et al. Human gut microbiota in obesity and after gastric bypass[J]. Proc Natl Acad Sci USA, 2009, 106(7): 2365-70.

[4] Costello EK, Lauber CL, Hamady M, et al. Bacterial community variation in human body habitats across space and time[J]. Science, 2009, 326(5960): 1694-7.

[5] Dethlefsen L, Huse S, Sogin ML, et al. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing[J]. PLoS Biol, 2008: e280.

[6] Roesch LF, Fulthorpe RR, Riva A, et al. Pyrosequencing enumerates and contrasts soil microbial diversity[J]. ISME J, 2007, 1(4): 283-90.

[7] Fulthorpe RR, Roesch LF, Riva A, et al. Distantly sampled soils carry few species in common[J]. ISME J, 2008, 2(9): 901-10.

[8] Lauber CL, Hamady M, Knight R, et al. Soil pH as a predictor of soil bacterial community structure at the continental scale: a pyrosequencing-based assessment[J]. Appl Environ Microbiol, 2009, 75(15): 5111-20.

[9] Rideout JR, He Y, Navas-Molina JA, et al. Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences[J]. PeerJ, 2014, 2: e545.

[10] Desantis TZ, Hugenholtz P, Larsen N, et al. Greengenes, a Chimera-Checked 16S rRNA gene database and workbench compatible with ARB[J]. Appl Environ Microbiol, 2006, 72(7): 5069-72.

[11] Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities[J]. Appl Environ Microbiol, 2005, 71(12): 8228-35.

[12] Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data[J]. Nat Methods, 2010, 7(5): 335-6.

[13] Schloss PD, Westcott SL, Ryabin T, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities[J]. Appl Environ Microbiol, 2009, 75(23): 7537-41.

[14] Cole JR, Wang Q, Cardenas E, et al. The ribosomal database project: improved alignments and new tools for rRNA analysis[J]. Nucleic Acids Res, 2009, 37(Database issue): D141-5.

[15] Glass EM, Wilkening J, Wilke A, et al. Using the metagenomics RAST server(MG-RAST) for analyzing shotgun metagenomes[J]. Cold Spring Harb Protoc, 2010(1): pdb.prot5368.

[16] Shah N, Tang H, Doak TG, et al. Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics[J]. Pac Symp Biocomput, 2011: 165-76.

[17] Walker AW, Duncan SH, Louis P, et al. Phylogeny, culturing, and metagenomics of the human gut microbiota[J]. Trends Microbiol, 2014, 22(5): 267-74.

[18] Wu X, Ma C, Han L, et al. Molecular characterisation of the faecal microbiota in patients with type II diabetes[J]. Curr Microbiol, 2010, 61(1): 69-78.

[19] Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes[J]. Nature, 2012, 490(7418): 55-60.

[20] He Y, Caporaso JG, Jiang XT, et al. Stability of operational taxonomic units:an important but neglected property for analyzing microbial diversity[J]. 2015: in press.

[21] Shanahan F, Quigley EM. Manipulation of the microbiota for treatment of IBS and IBD-challenges and controversies[J]. Gastroenterology, 2014, 146(6): 1554-63.

[22] Klatt NR, Funderburg NT, Brenchley JM. Microbial translocation, immune activation, and HIV disease[J]. Trends Microbiol, 2013, 21(1): 6-13.

[23] Gevers D, Kugathasan S, Denson LA, et al. The Treatment-Naive microbiome in New-Onset crohn's disease[J]. Cell Host Microbe, 2014, 15(3): 382-92.

(编辑:吴锦雅)